PROCEEDINGS *of the* FIFTH BERKELEY SYMPOSIUM ON MATHEMATICAL STATISTICS AND PROBABILITY

Held at the Statistical Laboratory
University of California
June 21–July 18, 1965
and
December 27, 1965–January 7, 1966

with the support of
University of California
National Science Foundation
National Institutes of Health
Air Force Office of Scientific Research
Army Research Office
Office of Naval Research

VOLUME IV

BIOLOGY AND PROBLEMS OF HEALTH

EDITED BY LUCIEN M. LE CAM
AND JERZY NEYMAN

UNIVERSITY OF CALIFORNIA PRESS
BERKELEY AND LOS ANGELES
1967

FREQUENCY DECISION THEORETICAL APPROACH TO AUTOMATED MEDICAL DIAGNOSIS

LEONARD RUBIN, MORRIS F. COLLEN
THE PERMANENTE MEDICAL GROUP, OAKLAND
and
GEORGE E. GOLDMAN
UNIVERSITY OF CALIFORNIA, BERKELEY

1. Introduction

The recent availability of relatively large memory capabilities associated with high speed computers has given hope to the possibility of combating some of the obvious deleterious results of disparate growth rates of medical knowledge, numbers of physicians, and numbers of patients. The purpose of this investigation is to establish a method for arriving at a preliminary medical diagnosis by the use of computer techniques before the patient is seen by the physician. It appears, however, that the serendipitous result of discovering more insight into the diagnostic process may be of greater medical significance than the proposed reason for the study.

2. Logic of diagnosis

In the past few years many investigators, employing a variety of approaches, have directed their efforts towards establishing a model for computer assisted medical diagnosis [1] to [7]. A frequent attempt at automating the medical diagnosis process consists of analyzing and then simulating the physician's diagnostic procedure. One problem with the attempt at simulation is apparent when one considers that whatever method the physician uses (heuristic, intuitive, probabilistic, deterministic, and so forth) it has the fault that if a disease is not considered at any stage of the process, that disease cannot be diagnosed. It is not likely that any physician ever considers *all* diseases in making a diagnosis. A series of two decision determinations encompassing all the diseases one wishes to consider obviates this problem.

Analyses of the currently employed physician diagnostic procedures reveals some interesting findings. There are inconsistencies in the current concepts of the diagnostic procedure. However, before these can be discussed it would be

Supported in part by a Community Health Services Project Grant from the U. S. P. H. S.

well to define some terms. The word characteristic will be used here to denote any information about a patient. Characteristics consist of several types of information among which are included identifying data (for example, age, sex), family history, symptoms, signs, test results, and so forth. Family history refers to any information concerning the patient's family. A symptom is any complaint which a patient makes about his condition. A sign is an objective abnormality determined by the examiner and detected by any of his five senses. A test is an examination designed to determine or measure a physiological, anatomical, or other variable. A disease is a specific combination of abnormal characteristics defined above.

It is generally assumed that symptoms are the results of one's having a specific illness. That is, one becomes afflicted with a disease and certain symptoms arise. This concept is supported by the existence of the term "asymptomatic disease" implying a state in which the disease is so early in its development that symptoms have not as yet started. In fact, it is one of the hopes of preventive medicine to be able to diagnose all diseases in the "presymptomatic stage." This would not only spare the patient the discomfort of the symptoms, but also permit treatment of the condition in its early stages. On the other hand, it is not always inferred that asymptomatic disease is necessarily early, or symptomatic disease necessarily late. Gout, for example, may herald its early appearance by very severe symptoms. Duodenal ulcer may develop late complications of hemorrhage and perforation without prior symptoms.

Another tradition of diagnosis is that ideally all symptoms should be independent of each other. This is an attempt to avoid redundancy, gain mathematical simplicity, and is based upon the intuitive feeling that independent symptoms intrinsically contain more information.

As noted above, a characteristic is any information concerning an individual. Thus, it may be any variable which can be associated with an individual, such as height, age, a blood chemistry result, and so forth. It also may be the result of a disease, such as pain or swelling, or partly be responsible for the onset of the disease, such as the fact that the individual ingested a toxic substance. Characteristics associated with a disease may be independent of each other within the nondiseased population, and yet be dependent on each other within the diseased group.

In medical diagnosis the kinds of characteristics that are actually used have no single logical relationship to each other and have no specific cause-effect relationship to the disease itself. Characteristics often are used without any idea of whether they precede or follow the disease. Some characteristics will be prerequisites for considering certain diagnostic categories. An example of this kind of characteristic is the necessity of being female to have the diagnosis of pregnancy. In this case, the characteristic does not derive from the disease and the disease does not necessarily derive from having the characteristic.

A more complex example demonstrating many of these problems is given by the clinical entity of diabetes mellitus. This is an example in which attempts to diagnose the disease at an earlier stage were partially responsible for a major disruption of the definition of the disease itself. The diagnosis "diabetes mellitus" in itself is almost meaningless today without further descriptive amplification. It is not easy to try to describe the complex relationships which exist between the characteristics of maternal diabetes, paternal diabetes, hyperglycemia, renal threshold, thirst, polyuria and diabetic retinopathy. Almost every degree of dependence and cause-effect relationship is represented. Some relationships have been reversed in the past few decades because of advancing medical knowledge.

Relationships between certain characteristics may be measured with great degrees of accuracy and specific correlations be given. However, for the vast majority of characteristics this cannot be done. Even if a relationship can be established, it is often not clear as to which is the causative agent and which is the resultant. It would be advantageous to use a diagnostic method which does not require that the relationship between the characteristics be of any specific kind, that is, be neither dependent, independent, causative nor resultant.

An estimate of the joint probability distribution of characteristics, that is, the probability of occurrence of characteristics in relationship to each other, may be obtained in two ways. The first method involves the use of existing medical knowledge and theory. Unfortunately, there is a serious deficiency in these areas and what does exist is subject to severe change, sometimes amounting to a complete reversal. The second method uses data collection and usually requires a prospective study because of the nature of present medical records. Retrospective studies for the collection of this kind of data reveal many inadequacies in records, as well as a great variety of descriptive phrases requiring interpretive judgments by the investigator.

The utilization of published medical information introduces a variety of problems. It would seem that the probability of a common symptom such as cough, in a common illness such as pneumonia, would be simple to determine. In medical books and journals, attempts at quantification usually will be couched in such terms as "not infrequently," "usually," "often," "rarely," "occasionally," "it has also been reported that," and so forth. Furthermore, these semiquantitative phrases apply to individual characteristics, not to the joint probability distribution of characteristics. Articles will be found dealing with specific numbers of cases of various illnesses and the probabilities of occurrence of individual characteristics. These are almost always derived from review of clinical records containing fragmentary data.

In our early work we went through a series of stages using what we thought were probably independent characteristics based on current medical theory. We then set up committees of experts who were to estimate, from their experience, the probability of occurrence of these "independent" characteristics. Disagreements usually resulted in compromise values of the probabilities of individual characteristics in the disease group. Consider the problem of trying to establish agreement as to the probabilities of combinations of characteristics. Soon, the conclusion was reached that in order to get meaningful data, prospective studies involving the simultaneous collection of characteristics in large numbers of patients was required.

An important attribute of characteristics which is not always considered in automated diagnosis is that of time. Time relationship between different characteristics (completely aside from the problem of causality) is important along with the sequence in which different characteristics appear. Another consideration on which attention must be focused is the time interval of recurrence of the same characteristic and its duration in time. Still another temporal aspect of characteristics which must be considered is that of the relationship of therapy to the modification of the characteristic. The physician uses all of these time relationships in his diagnostic procedure. Again, these facts are not easily obtained from the literature or from experienced physicians. Prospective studies may be adjusted so as to include one or two of the time considerations in the data collection itself, as will be noted in tables I and II. However, it is not feasible to qualify fully the characteristics in regard to time because of the huge increase in the sample space which is then necessary.

Another kind of information used in arriving at a medical diagnosis is the probability of the existence of the disease in the population under consideration, usually referred to as the prevalence of a disease. If "population under consideration" is to be defined as the geographical area, how big should this area be? The prevalence of a disease in a city may be totally different from that of the state. San Joaquin Valley fever is such an example. Available statistics are inadequate because of the methods of collection of these statistics.

The milieu of the patient is ever changing. In the midst of an epidemic should a patient telephone his physician and report the development of symptoms compatible with that of the epidemic disease, the diagnosis would most likely be that of the epidemic. However, should the patient go on to relate that several members of his family are currently afflicted with another contagious disease, the physician would be hard pressed to decide upon a diagnosis. What *a priori* prevalence rate for a disease should be used? That of the state, city, block, factory, family? At the start of an epidemic this changes by the hour. For some diseases it changes with the seasons; for some it changes with the landing of an airplane at the local airport. When a situation arises in which information of this sort is very important, it can be incorporated into the characteristic complex in this method. For example, residence or visit in the San Joaquin Valley would be one of the characteristics used in consideration of San Joaquin Valley Fever.

Up to this point consideration has been given primarily to the characteristics of the patient. Relationships between the diseases to be diagnosed must also be considered. Unfortunately, extraneous diseases can act on a patient in such a way as to interfere with a particular disease process and its manifestations. The coexistence of other diseases can severely modify the characteristics produced by each single disease. As compared with methods which use mutually exclusive diseases, a sequential two decision approach to the problem has advantages in this regard, because at each step the question is merely: does or does not the patient have the disease or disease combinations under consideration *regardless* of what other disease he may have. In the proposed method this is inherently true because of the manner in which the test is established.

In any system in which the probabilities of the existence of mutually exclusive and exhaustive diseases are reported, the probabilities of all states considered, diseased and nondiseased, must sum to unity and one is therefore dealing with a closed system. Should a patient have a disease outside the closed system as constructed, the reliability of the diagnosis is decreased. This may obtain by not having a disease within the specified system or by having a nonspecified combination of the diseases. It is important to avoid the use of a system in which diseases are mutually exclusive unless one is prepared to handle every combination of all known diseases. In a sequential two decision problem, the diseases considered at a particular two decision level may have any or no relationship to diseases considered at other levels.

In view of the foregoing, the following criteria were set in deciding upon a method for approaching automated medical diagnosis:

(1) any disease, disorder, or category should be eligible for inclusion either alone or in any combination with any other disease, disorder or category;

(2) any characteristic should be usable regardless of apparent relationships either to other characteristics or to the disease itself;

(3) disease prevalence data should not be used as such but information contained in such figures, when it is of sufficient significance, should be included as specific characteristics;

(4) to avoid discarding useful characteristics on the basis of currently held but invalid medical theory, the automated method should itself select the most useful characteristics.

3. Mathematical model for symptom selection and diagnosis

Responses to 50 dichotomous questions were collected from known "normal" persons and patients with specific disease entities. The form of the questions used was such that it was expected that diseased patients would answer affirmatively and "normals" would answer negatively. Data reduction to 12 of these 50 dichotomous questions was accomplished on the basis of single characteristic marginal attributes. The probabilities of affirmative responses for each single characteristic in both the diseased and "normal" groups were determined. The ratios of these probabilities, diseased to "normal," was then calculated and ranked in order of magnitude. The 12 highest ratios, providing that an affirmative response in the diseased group of at least 10 per cent was obtained, indicated the 12 characteristics to be selected. The tabulation was further examined to be certain that no characteristic had a significantly greater negative response rate in the diseased group.

The necessity for limiting the number of characteristics to a small set, such as 12, at this time is apparent from a consideration of several factors. The number of possible responses to 12 dichotomous questions is 2^{12} or 4096. The number of patients in the study is about 12,000 for the "normals" and approximately 500 for the largest group of diseased patients. The number of patients per possible characteristic response is not sufficient to fill an adequate number of cells when using more than 12 characteristics. Additional limitations arise from the memory and speed capabilities of even large modern computers.

At this point the critical regions established by the Neyman-Pearson method of testing hypotheses [8] were employed not only for elimination of less valuable characteristics, but also to establish those sets of responses that will be used to classify an individual as having the particular disease under consideration. Having reduced the number of characteristics to 12 for each disease, the remainder of the procedure is accomplished by the computer, using the estimated errors to eliminate the less valuable characteristics.

The basis of the method consists of calculating the ratio θ_{Si}, which is the proportion of patients with the disease under consideration and with a specific set of characteristic responses P^D_{Si}, to the proportion of individuals without this disease and who have the same set of characteristic responses P^N_{Si}. This is done for all sets of characteristic responses and results in a ratio for each characteristic response set,

$$\theta_{Si} = \frac{P^D_{Si}}{P^N_{Si}}. \tag{3.1}$$

These ratios are then assembled in ascending order of magnitude and the characteristic response set associated with each ratio is arranged in the same order. It will be found that most individuals with characteristic response sets near the top of the order (low ratios) are nondiseased (with reference to the disease in question), and those individuals with characteristic response sets near the bottom of the array (higher ratios) have the disease. A criterion can then be established to divide the array of θ into two regions, nondiseased and diseased, depending on the errors which can be tolerated. The criterion is a numerical value of θ and the characteristic response patterns above this value are classified as "nondiseased" and the characteristic response patterns below this value are classified as "diseased." Since the hypothesis we are testing is that an individual is diseased, a criterion near the top of the array will have a lower probability of "error of the first kind" or α, and a higher probability of "error of the second kind." A criterion near the top of the array usually will in effect classify more "normals" as diseased, but will not usually classify many diseased as "normals" (lower rejection of the true hypothesis). As the criterion is moved down the array the reverse obtains. The method insures that having selected a required estimated probability of error of the first kind, perhaps because of the nature of the disease under consideration, the estimated power or estimated specificity of the test is maximized. Conversely, should a specific power or specificity be required, the sensitivity (which is equal to $1 - \alpha$) will have to be maximized.

The estimated probability of error is the empirical probability found in the array. As the sample size increases these estimates approach the true error. For the remainder of this paper "error" refers to the estimated error unless otherwise qualified.

Further data reduction, employing the above described method, to a single characteristic, was accomplished by a step down procedure in which characteristics were eliminated one at a time. A step down procedure was employed because it seemed the most direct way to use the unknown relationships between the characteristics to a maximal practical limit. This method does not use all the information present and therefore does not necessarily select the best characteristics even assuming the estimated probabilities are the real probabilities. To do so would require the examination of all combinations of all sets of characteristics and their responses which at this time is not feasible. The criterion used for the elimination of these characteristics, one at a time, is a linear function of the two kinds of estimated errors,

$$\text{criterion} = a(1 - \alpha) + b(\text{power}). \tag{3.2}$$

At each step (reducing n characteristics to $n - 1$ characteristics) the set of $n - 1$ characteristics is selected which maximizes the criterion.

Using these $n - 1$ characteristics as the new base set, the process is repeated resulting in a selection of $n - 2$ characteristics. In this manner characteristics are selected for each disease under consideration. In the diseases considered here a and b were temporarily set at unity. For other conditions, such as tuberculosis, a would probably be set several magnitudes higher than b. This may be established for any disease in an optional manner.

Having now completed the ranking of individual characteristics with respect to value in discriminating the disease in question, attention must be directed to the sensitivity and specificity required by the medical considerations of that disease. As will be shown, such a consideration will indicate the minimum number of characteristics which must be used for that disease. Having selected the n "best" characteristics for diagnosing the disease in question, the array of ordered θ is used to establish which sets of characteristic responses will differentiate the diseased and "normal" groups. The sample space, that is, all possible sets of characteristic responses, will be divided into two regions by a criterion established by the necessary sensitivity and specificity. The "negative region" occupies the part of the array at or above the criterion value and includes all sets of characteristic responses typifying those individuals who are likely to be free of the disease in question. The "positive region" is that part of the array below the criterion value and includes those sets of characteristic responses more likely to be found in those individuals who have the disease in question.

Having established the characteristic response patterns in each region it is henceforth necessary to determine only which characteristic response pattern an individual has and then determine in which region of the sample space it lies.

4. Example of the method

A number of considerations were involved in selecting diseases with which to establish a test for this method. Because it was planned to employ a relatively large number of characteristics it was apparent that many individual patients would be necessary, thus requiring diseases with high prevalence rates.

Diseases in which definite diagnoses could be confirmed by some objective method would be desirable. Preferably, characteristic responses should be obtained prior to the establishment of the diagnosis, so that the patients would not be influenced by the fact that they had just been informed of the diagnosis. It was further decided to select two diseases which were difficult to differentially diagnose clinically to test the method under very difficult circumstances. All patients who were about to have X-ray examinations of the upper gastrointestinal tract were requested to fill out a questionnaire prior to having the X-ray examination done. These questionnaires consisted of 50 questions selected from several hundred used in our Multiphasic Health Checkup [9], which, on the basis of medical theory and knowledge, would seem to be those containing the most information of use in diagnosing diseases of the gastrointestinal tract.

The questionnaires were then collected and set aside until the reports of the X-ray examinations were completed. All X-ray reports indicating the presence of a hiatus hernia or duodenal ulcer (in any stage of activity) were then collated with the questionnaire for that patient. Of necessity this procedure introduced a number of variables which could not be avoided without decimating the data available for the study. Among these variables are the use of different radiologists, the use of some individuals who knew they had the disease in question from past examinations, variable time of onset of symptoms relative to time of completion of the questionnaire, possible onset of therapy for the disease based on clinical grounds before the X-ray was done, and so forth. All of these variables tend to decrease the ease of diagnosing the disease in question.

Because of the nature of medical examinations today, it is not possible to obtain a large group of individuals who are known to be free of the above named diseases and who are also free of other closely associated abnormalities. These X-rays are not part of a routine examination done on all individuals. Patients are not subjected to these radiological examinations unless there is some deviation from normal suggesting that the examination be performed. Accordingly, it was decided to use as the "nondiseased" group all individuals reporting for the Multiphasic Health Checkup during a specified period of time and in whom, as a matter of course, all of the identical questions would have been asked. It was felt that the prevalence of the conditions to be studied was such that an error of not more than perhaps five per cent was introduced, and this was always in the direction of having some diseased patients in the "normal" group. This "nondiseased" group of 12,262 individuals will henceforth be referred to as "normals."

In order to determine how patients with some other diseases might be divided in the sample space, several diagnostic categories were selected from those diagnoses made by physicians following the Multiphasic Health Checkup examinations. These individuals had answered the same 50 questions which had been asked of the patients subjected to the radiological examination. These diseases were selected in such a manner that some were nongastrointestinal diseases which might ordinarily be misdiagnosed as the gastrointestinal diseases under study. Among these were angina pectoris and ischemic heart disease. Other diseases were selected because in no way should there be any confusion with the gastrointestinal diseases. In this group were benign prostatic hypertrophy and bronchial asthma.

By applying the procedure described above for selection of characteristics most useful in differentiating diseased from "normal" patients by use of marginal attributes, the 50 questions were reduced to 12. Those found to be most valuable for diagnosis of hiatus hernia are listed in decreasing order of value in table I. Those found to be most valuable for diagnosis of duodenal ulcer are listed in decreasing order of value in table II. Such questions as "Do you have a duodenal ulcer?" were eliminated from the study for obvious reasons, although physicians do find these questions of great assistance in diagnosis. It will be noted that there are many questions common to both lists.

TABLE I

QUESTIONS USED FOR DIAGNOSIS OF HIATUS HERNIA

In decreasing order of value according to the criterion established for this study. (Questionnaire number precedes each question.)

31. HAVE YOU *often*, IN THE PAST *6 months*, HAD HEARTBURN, INDIGESTION OR PAIN IN THE *upper* PART OF YOUR STOMACH OR BELLY (*above* the navel or belly button)?
38. HAVE YOU OFTEN, IN THE PAST *6 months*, HAD HEARTBURN, INDIGESTION OR PAIN IN YOUR STOMACH THAT AWAKENED YOU FROM SLEEP?
16. HAVE YOU *often*, IN THE PAST *6 months*, HAD TROUBLE SWALLOWING SOLID FOOD?
63. HAVE YOU *often*, IN THE PAST *6 months*, TAKEN LAXATIVES OR CATHARTICS?
28. HAVE YOU *often*, IN THE PAST *6 months*, HAD NAUSEA (sick to your stomach) OR VOMITING (throwing up)?
21. HAVE YOU *often*, IN THE PAST *year*, HAD PAIN OR PRESSURE OR A TIGHT FEELING IN YOUR CHEST WHEN YOU WERE SITTING STILL?
23. HAVE YOU, IN THE PAST *year*, HAD REPEATED PAIN OR PRESSURE OR A TIGHT FEELING IN YOUR CHEST AFTER A BIG MEAL?
19. HAVE YOU *often*, IN THE PAST *year*, HAD PAINS IN THE SIDES OF YOUR CHEST OR ON ONE SIDE ONLY (and *not* in the middle of your chest)?
27. HAVE YOU *often*, IN THE PAST *year*, HAD PAIN IN YOUR CHEST THAT LASTED *more* THAN TEN MINUTES?
37. HAVE YOU *often*, IN THE PAST *six months*, HAD INDIGESTION, HEARTBURN OR PAIN IN YOUR STOMACH BROUGHT ON OR MADE WORSE BY BENDING OVER OR LYING DOWN?
40. HAVE YOU *often*, IN THE PAST *6 months*, HAD PAIN IN YOUR STOMACH THAT WAS NOT HELPED BY MEDICINES?
24. HAVE YOU, AT ANY TIME IN THE PAST *year*, HAD ATTACKS OR EPISODES OF PAIN OR PRESSURE OR TIGHT FEELING IN YOUR CHEST THAT AWAKENED YOU FROM SLEEP?

TABLE II

QUESTIONS USED FOR DIAGNOSIS OF DUODENAL ULCER

In decreasing order of value according to the criterion established for this study. (Questionnaire number precedes each question.)

31. HAVE YOU *often*, IN THE PAST *6 months*, HAD HEARTBURN, INDIGESTION OR PAIN IN THE *upper* PART OF YOUR STOMACH OR BELLY (*above* the navel or belly button)?
38. HAVE YOU *often*, IN THE PAST *6 months*, HAD HEARTBURN, INDIGESTION OR PAIN IN YOUR STOMACH THAT AWAKENED YOU FROM SLEEP?
60. HAVE YOU *often*, IN THE PAST *6 months*, TAKEN MEDICINE FOR YOUR STOMACH OR DIGESTION?
28. HAVE YOU *often*, IN THE PAST *6 months*, HAD NAUSEA (sick to your stomach) OR VOMITING (throwing up)?
41. HAVE YOU *often*, IN THE PAST *6 months*, HAD HEARTBURN, INDIGESTION OR PAIN IN YOUR STOMACH THAT WAS RELIEVED OR HELPED BY EATING?
30. FOR THE PAST *month* HAVE YOU HAD A POOR APPETITE FOR MOST OF THE TIME?
34. HAVE YOU *often*, IN THE PAST *6 months*, HAD PAIN IN THE LEFT SIDE OF YOUR STOMACH OR BELLY?
40. HAVE YOU *often*, IN THE PAST *6 months*, HAD PAIN IN YOUR STOMACH THAT WAS NOT HELPED BY MEDICINES?
33. HAVE YOU *often*, IN THE PAST *6 months*, HAD PAIN IN THE RIGHT SIDE OF YOUR STOMACH OR BELLY?
51. HAVE YOUR BOWEL MOVEMENTS *often* BEEN MIXED WITH MUCUS OR SLIMY MATTER IN THE PAST *6 months?*
64. HAVE YOU LOST MORE THAN 10 POUNDS, IN THE PAST *6 months*, WITHOUT TRYING?
54. HAVE YOU, IN THE PAST YEAR, HAD ANY SOFT BLACK (as tar) BOWEL MOVEMENTS (when *not* taking iron medicine or vitamins with minerals)?

The 12 questions were reduced in number to a single characteristic, one at a time by the iterative procedure described above. As a result of this procedure an array of sets of characteristic responses and the associated θ was obtained for the "best" characteristics at each level of this step down procedure. Examples of such arrays are shown in tables III and IV.

TABLE III

COMBINATIONS OF SIX CHARACTERISTICS FOR DIAGNOSIS OF DUODENAL ULCER AT SPECIFIED SENSITIVITIES AND SPECIFICITIES

Characteristics are questions number 31, 38, 60, 28, 41, 30 of table II.

Characteristic Set	Duodenal Ulcer Patients		"Normals"		θ	Estimated Sensitivity Per Cent	Estimated Specificity Per Cent
	No.	Per Cent	No.	Per Cent			
23	0	0	1	0.000	0.00	100	1
52	0	0	2	0.000			
54	0	0	3	0.000			
51	0	0	4	0.000			
20	0	0	11	0.001			
25	0	0	27	0.2			
34	0	0	28	0.2			
3	0	0	80	0.7			
0	29	8.7	8582	70.0	0.12	91	71
2	1	0.3	231	1.9	0.16	91	73
16	4	1.2	240	2.0	0.61	90	75
4	2	0.6	114	0.9	0.64	89	76

TABLE III (Continued)

Characteristic	Duodenal Ulcer Patients		"Normals"		θ	Estimated Sensitivity	Estimated Specificity
Set	No.	Per Cent	No.	Per Cent		Per Cent	Per Cent
8	12	3.6	538	4.4	0.82	86	80
32	6	1.8	212	1.7	1.04	84	82
1	12	3.6	388	3.2	1.14	80	85
9	9	2.7	267	2.2	1.24	78	87
42	1	0.3	28	0.2	1.31	77	88
5	3	0.9	68	0.6	1.62	76	88
48	2	0.6	42	0.3	1.75	76	89
7	2	0.6	34	0.3	2.16	75	89
39	1	0.3	17	0.1			
36	1	0.3	16	0.1	2.30	75	89
10	11	3.3	161	1.3	2.51	71	91
12	7	2.1	94	0.8	2.73	69	91
18	2	0.6	26	0.2	2.82	69	91
41	4	1.2	49	0.4	3.00	67	92
40	6	1.8	70	0.6	3.15	66	92
33	3	0.9	34	0.3	3.24	65	93
11	15	4.5	149	1.2	3.70	60	94
13	14	4.2	139	1.1	3.70	56	95
17	2	0.6	19	0.2	3.87	55	95
6	4	1.2	34	0.3	4.32	54	96
24	5	1.5	40	0.3	4.59	52	96
26	1	0.3	8	0.1			
21	1	0.3	6	0.0	6.12	52	96
38	1	0.3	6	0.0			
50	1	0.3	6	0.0			
19	1	0.3	5	0.0	7.34	51	96
55	1	0.3	5	0.0			
47	12	3.6	57	0.5	7.73	47	97
35	5	1.5	23	0.2	7.98	46	97
57	2	0.6	9	0.1	8.16	45	97
29	3	0.9	13	0.1	8.47	44	97
14	13	3.9	56	0.5	8.52	40	97
43	9	2.7	35	0.3	9.44	38	98
15	33	9.9	113	0.9	10.72	28	99
37	3	0.9	10	0.1	11.01	27	99
56	4	1.2	13	0.1	11.30	26	99
27	2	0.6	6	0.0	12.24	25	99
62	1	0.3	3	0.0			
45	14	4.2	38	0.3	13.53	21	99
30	2	0.6	5	0.0	14.69	19	99
60	2	0.6	5	0.0			
46	7	2.1	16	0.1	16.06	17	99
61	6	1.8	13	0.1	16.94	16	99
58	2	0.6	4	0.0	18.36	15	99
44	9	2.7	17	0.1	19.44	12	99
31	7	2.1	11	0.1	23.36	10	99
63	13	3.9	17	0.1	28.07	6	99
22	1	0.3	1	0.0	36.71	5	99
53	3	0.9	3	0.0			
28	3	0.9	2	0.0	55.07	4	99
49	5	1.5	3	0.0	61.19	3	99
59	9	2.7	5	0.0	66.08	0	100

Table III contains the array for the "best" six characteristics for distinguishing patients with duodenal ulcer from "normal" individuals. The number of patients with duodenal ulcer is 334. With six characteristics there are 64 possible sets of characteristics and in this case all 64 were represented by at least one individual from at least one of the groups.

TABLE IV

COMBINATIONS OF FOUR CHARACTERISTICS FOR HIATUS HERNIA DIAGNOSIS AT SPECIFIED SENSITIVITIES AND SPECIFICITIES

Characteristics are questions number 31, 38, 16, 63 of table I.

Characteristic Set	Hiatus Hernia Patients No.	Per Cent	"Normals" No.	Per Cent	θ	Estimated Sensitivity (per cent true +)	Estimated Specificity (per cent true −)
0	88	16.8	9119	74.4	0.23	83	74
1	12	02.3	696	05.7	0.40	81	80
2	19	03.6	258	02.1	1.7	77	82
8	8	01.5	91	00.8	1.9	76	83
9	1	00.2	17	00.1	2.0	76	83
4	104	19.9	1189	09.7	2.1	56	93
12	8	01.5	33	00.3	5.0	52	94
3	13	02.5	59	00.5			
5	44	08.4	172	01.5	5.6	43	95
6	132	25.2	450	03.7	6.8	18	99
15	10	01.9	19	00.2	9.5	16	99
7	44	08.4	94	00.8	10.5	08	99
13	9	01.7	15	00.1	17.0	06	99
14	27	05.2	36	00.3	17.3	01	99
11	2	00.4	7	00.0	—	01	99
10	3	00.6	7	00.0	—	00	100
Total	524		12262				

Table IV contains the array for the "best" four characteristics found to distinguish 524 patients with hiatus hernia from the 12,262 "normals." Again, some of these "normals" have hiatus hernia in accordance with the prevalence of hiatus hernia in this population. Again, all 16 possible sets of characteristics are represented.

It will be noted that if the array is divided into an upper and lower region at any optional sensitivity, the specificity is established. The higher the sensitivity desired the lower the specificity. The method insures that at the selected sensitivity the estimated specificity is maximized.

It will be noted in table III that if the criterion used for selection of duodenal ulcer cases was set at a sensitivity of 86 per cent, the specificity or power would be 80 per cent. Thus, all individuals who had characteristic response sets associated with θ values of 0.82 or less would be classified as being "normal" since they belong to the upper "negative region" and those with characteristic response sets associated with a θ greater than 0.82 would lie in the "positive region" and be classified as having duodenal ulcer.

Although 12,262 individuals comprise the "normal" group, and there are only 64 cells to fill in table III, it will be noted that only one "normal" person had characteristic response set 23. Clearly, this is a weak assignment of this cell to the negative region because if only one duodenal ulcer patient were to have this characteristic response set the θ would have been 36.7 and this cell would have been categorized as belonging to the positive region. Similarly, there are many other cells assigned to either the positive or negative region on the basis of the response of a single individual. In those arrays created by more than six characteristics there are cells which are completely unspecified and cannot be identified as belonging to either region. The more characteristics used the greater the number of unspecified cells.

It will be noted in table IV that with only four characteristics and 16 cells there is only a single characteristic response set that was assigned to a region based on the response of a single individual. Specifically, characteristic response set 9 would have been in the negative region had it not been for the one hiatus hernia patient with this characteristic response set.

Another noteworthy comparison of these two arrays reveals that the characteristic response set in which all individuals respond negatively to all questions, set 0, lies midway in the array when many characteristics are used and advances to the top of the array as the number of characteristics used is decreased. When this occurs a maximum value has been set for the sensitivity which may be selected because it is not medically desirable to randomly divide patients who have identical characteristic responses into diseased and nondiseased categories. It indicates that even if one should classify as diseased any individual who answered any single characteristic positively, some diseased patients would still be classified as nondiseased.

From the foregoing it is apparent that there are patients without any of these symptoms who were sent for the gastrointestinal X-ray examination. There were in fact 59 hiatus hernia patients who answered negatively to all of the "best" 11 of the 12 questions. Analysis of the clinical records of these individuals revealed that patients may be sent for X-ray examinations which reveals the presence of hiatus hernia or duodenal ulcer even though they have no symptoms at all. Physicians have known that both of these conditions may exist in the absence of any symptoms. In this study, individuals without symptoms were sent for X-rays for a variety of reasons among which were: (1) anemia of unknown cause, the X-ray being done to discover possibly some silent lesion; (2) X-ray required for some highly sensitive job requirement such as airline pilot; (3) dermatological condition present which is often associated with radiological findings on gastrointestinal X-ray examination; (4) intermittent symptoms lasting a very short time and not present several days later at the time of the X-ray; and (5) pernicious anemia in which it is a frequent procedure to get periodic X-ray examinations because of the increased prevalence of neoplasm. It is apparent that in view of the existence of these diseases (hiatus hernia and duodenal ulcer) in the absence of symptoms, no method of diagnosis depending on symptoms will be 100 per cent successful. Similarly, because of the existence of individuals with the disease in the "normal" group, a specificity of 100 per cent would not be expected.

In order to test the arrays with additional cases of hiatus hernia and duodenal ulcer to determine the actual error of the method, 50 cases of each disease, selected in the same manner as the previous groups but not used in establishing the array, were examined. The characteristic response patterns were determined for all individuals using the single "best" characteristic and combinations of as many as the "best" eight characteristics. The criterion was the maximum sum of the sensitivity and specificity. This criterion had been used in the selection of the "best" characteristics to prevent the sensitivity or specificity from extending outside the bounds of clinical usefulness. Had the criterion been set at an arbitrary sensitivity, for example, characteristics could have been selected when the specificity was either too high or too low to be useful. However, any criterion might have been chosen depending upon the requirements demanded of the disease in question. The results of this test are shown in tables V and VI. In addition, these tables show the result of testing 579 additional Multiphasic Health Checkup (MHC) individuals. These are not a subset of the 12,262 individuals whose responses were used to establish the array. Also included in these tables are the results with the other diseases tested. In addition, the hiatus

TABLE V

ESTIMATED AND ACTUAL SENSITIVITY AND SPECIFICITY IN DIAGNOSIS OF HIATUS HERNIA
Criterion for positive region is the maximum sum of estimated sensitivity and specificity.

	Sensitivity (per cent)		Specificity (per cent)						
				Actual					
Quantity of Symptoms Used	Estimate	Actual 50 Cases	Estimate	579 Routine Patients (MHC)	384 Duod. Ulcer Cases	96 Asthma Cases	114 Angina Pectoris Cases	168 Ischaemic Heart Disease Cases	238 Benign Prostatic Hypertrophy Cases
1	72.1	68.0	83.6	83.9	27.1	84.4	83.3	82.7	89.5
2	79.2	80.0	80.9	81.0	21.1	76.0	81.6	79.2	87.0
3	80.9	82.0	80.0	80.1	19.8	74.0	79.8	78.0	85.7
4	80.9	82.0	80.1	80.1	19.7	78.9	79.6	78.0	86.7
5	81.9	82.0	79.8	79.5	18.9	74.0	79.8	78.0	85.3
6	81.3	82.0	81.2	80.1	20.2	81.3	79.8	80.4	86.6
7	81.7	80.0	81.6	80.0	22.1	81.2	78.9	79.2	87.0
8	80.2	66.0	84.2	82.6	22.7	77.1	82.5	79.2	87.0

hernia patients were tested with the duodenal ulcer symptoms and the duodenal ulcer patients were tested with the hiatus hernia symptoms.

TABLE VI

Estimated and Actual Sensitivity and Specificity in Diagnosis of Duodenal Ulcer

Criterion for positive region is the maximum sum of estimated sensitivity and specificity.

	Sensitivity (per cent)		Specificity (per cent)						
				Actual					
Quantity of Symptoms Used	Estimate	Actual 50 Cases	Estimate	579 Routine Patients (MHC)	574 Hiatus Hernia Cases	96 Asthma Cases	114 Angina Pectoris Cases	168 Ischaemic Heart Disease Cases	238 Benign Prostatic Hypertrophy Cases
1	71.4	84.0	83.5	83.9	28.2	84.4	83.3	82.7	89.5
2	78.1	84.0	80.8	81.0	20.7	76.0	81.6	79.2	87.0
3	86.5	86.0	76.4	76.3	17.9	69.8	76.3	70.2	81.9
4	89.2	92.0	74.0	73.4	15.7	64.6	75.4	67.9	80.7
5	88.3	92.0	75.8	75.1	17.9	69.8	76.3	70.8	82.8
6	85.6	84.0	80.4	80.0	28.1	71.8	83.3	76.8	86.2
7	86.8	80.0	80.9	80.4	27.6	73.2	95.9	77.3	85.0
8	85.6	70.0	84.0	84.0	36.9	77.1	88.6	79.9	86.1

5. Discussion

Examination of the data in tables V and VI reveals that the gastrointestinal diseases, hiatus hernia and duodenal ulcer, are sufficiently like each other that they cannot be separated when only a few characteristics are employed. Some differentiation is apparent in table VIII when using five or more duodenal ulcer characteristic responses of the hiatus hernia patients.

It is important to note that the sensitivities and specificities as shown in tables III and IV are discrete points. For clinical purposes of testing on individual patients, interpolation between these points should not be done, since one would then be forced to arbitrarily and randomly divide a group of patients with identical responses into two divisions as required by the interpolation. Patients with identical characteristics would then be found in both the positive and negative regions. This means that should a sensitivity of 80 per cent be desired in the case of hiatus hernia (table IV) one would be forced to make the clinical decision between 81 per cent and 77 per cent. However, for purposes of comparison of symptoms and methods, when one is not testing specific patients, interpolation does enable one variable to be held constant and allows for better examination of the behavior and characteristics of the method.

TABLE VII

ESTIMATED AND ACTUAL SENSITIVITY AND SPECIFICITY IN DIAGNOSIS OF HIATUS HERNIA
Criterion for positive region is the maximum sum of estimated sensitivity and specificity.
Estimated sensitivity fixed at 80 per cent.

	Sensitivity (per cent)		Specificity (per cent)						
				Actual					
Number Variables	Estimate	Actual 50 Cases	Estimate	579 Routine Patients (MHC)	384 Duod. Ulcer Cases	96 Asthma Cases	114 Angina Pectoris Cases	168 Ischaemic Heart Disease Cases	238 Benign Prostatic Hypertrophy Cases
1	80.0	77.0	60.5	61.0	19.7	61.4	60.6	60.2	65.0
2	80.0	80.8	77.8	77.9	20.4	73.0	78.4	76.2	83.6
3	80.0	80.9	80.5	80.6	20.7	75.1	80.7	78.7	86.4
4	80.0	79.7	80.6	80.5	20.7	80.7	80.0	78.5	87.3
5	80.0	78.4	81.4	80.4	21.0	81.4	81.4	80.1	87.5
6	80.0	79.7	82.1	80.6	20.5	81.6	81.8	81.2	88.1
7	80.0	79.7	82.8	81.8	22.7	82.6	81.8	81.2	88.5
8	80.0	66.0	84.3	82.7	22.7	77.1	82.5	79.2	87.1

The data of tables V and VI are shown in this latter form in tables VII and VIII. Here the sensitivities have been held at 80 per cent and the changes in specificity as the number of characteristics is increased is more apparent.

TABLE VIII

ESTIMATED AND ACTUAL SENSITIVITY AND SPECIFICITY IN DIAGNOSIS OF DUODENAL ULCER
Criterion for positive region is the maximum sum of estimated sensitivity and specificity.
Estimated sensitivity fixed at 80 per cent.

	Sensitivity (per cent)		Specificity (per cent)						
				Actual					
Number Variables	Estimate	Actual 50 Cases	Estimate	579 Routine Patients (MHC)	574 Hiatus Hernia Cases	96 Asthma Cases	114 Angina Pectoris Cases	168 Ischaemic Heart Disease Cases	238 Benign Prostatic Hypertrophy Cases
1	80.0	88.3	61.1	61.4	20.6	61.8	61.0	60.6	65.6
2	80.0	85.4	73.8	74.0	18.9	69.4	74.5	72.4	79.4
3	80.0	79.8	80.1	81.4	22.9	77.0	78.5	72.8	85.4
4	80.0	80.8	81.8	81.7	23.2	75.5	82.9	80.0	88.0
5	80.0	79.4	83.4	82.5	34.1	74.4	84.3	78.2	86.7
6	80.0	75.6	85.5	85.9	31.5	79.3	87.8	85.8	91.3
7	80.0	72.7	86.8	85.7	31.5	80.1	87.6	83.9	88.0
8	80.0	56.0	88.6	89.7	40.6	84.4	91.2	88.8	91.1

The data in these tables reveal that there is no difficulty obtaining acceptable specificities with either the hiatus hernia or duodenal ulcer symptoms when patients with diseases other than gastrointestinal diseases are used. While the characteristics employed here do not satisfactorily distinguish between the two gastrointestinal diseases it should be pointed out that the characteristics were not selected for this specific purpose. These characteristics were selected to give the "best" differentiation from the "normal" patients as previously described. Several analyses were made on symptoms selected specifically for the purpose of distinguishing these two diseases and some improvement in diagnostic ability was obtained. However, this is not a practical solution to this problem because it implies that the best characteristics would have to be sought to differentiate between all combinations of diseases. This could be used as an alternate method in selected situations in which the differential diagnosis could only be made by more hazardous medical methods, such as surgical exploration.

One of the prime problems still confronting any method of automated medical diagnosis involves the selection of characteristics to be used. This is of importance because there is not a wide latitude in the number of characteristics which may be used. Figures 1 and 2 show the increase in specificity obtained, with a

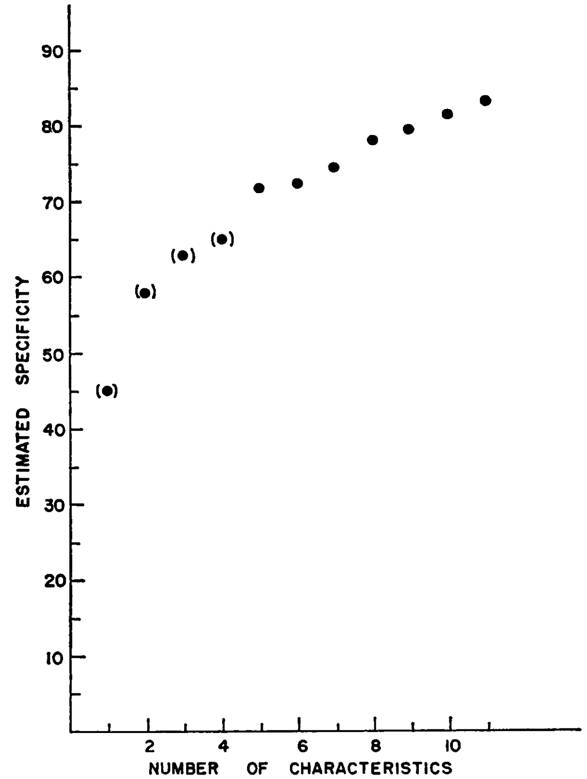

FIGURE 1

Hiatus hernia estimated sensitivity = 85 per cent.

fixed sensitivity, when increasing numbers of characteristics are employed. The points in brackets were extrapolated and indicate theoretical points which cannot be used without randomly dividing the patients at a particular θ level. Thus, at the sensitivity indicated, 85 per cent in figure 1, it is necessary to use at least five characteristics. From table IV it can be seen that sensitivities below 83 per cent may be achieved with the use of four characteristics. It will be noted from figures 1 and 2 that the increments in specificity are not large as the number of characteristics increases. It must be remembered also that the numbers of characteristic response sets increases logarithmically with the increase in numbers of characteristics. An increasingly large number of characteristic response sets become unidentified (with respect to belonging to the positive or negative region) as the number of characteristics is increased. This results in decreased performance as shown by the decrease in sensitivity in tables VII and VIII when the 50 additional cases are diagnosed by the use of eight rather than seven characteristics. In this study the positive region was established by the use of only 524 and 334 patients, respectively.

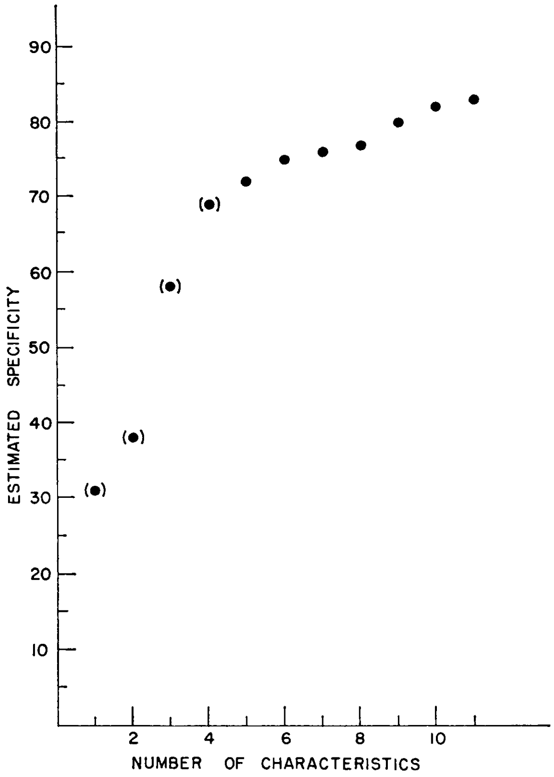

FIGURE 2

Duodenal ulcer estimated sensitivity = 90 per cent.

Because of these limitations on the number of characteristics used, it is extremely important to select the characteristics that carry the greatest amount of information. It is for this reason that it would be advantageous to start the automated procedure in the earliest stages of data reduction so that valuable characteristics are not overlooked because of current medical opinion. A number of approaches were attempted in this regard and the method selected and herein described was slightly better than other methods tried. Unfortunately, no properties of the marginal values of the individual characteristics were discovered which were clearly more useful than what was done in making the "best" selection. If such a property exists it will be a function of the proportions of affirmative responses of the diseased and nondiseased groups as well as the relationships between the characteristics.

6. Conclusions

The critical regions established by the Neyman-Pearson method of testing hypotheses does enable individuals to be classified on the basis of response to questions as diseased or nondiseased with medically acceptable levels of sensitivity and specificity. The use of too few characteristics results in estimated errors, and the use of too many characteristics results in actual errors, which in either instance are too great for clinical usefulness. A sequential two decision plan obviates some of the difficulties encountered in the medical diagnostic process.

The assistance of L. Davis in computer programming, A. Siegelaub in statistical analyses, J. Neyman and G. Dantzig in advice and counsel, and the assistance of the Computer Center Staff, University of California, Berkeley, is gratefully acknowledged.

REFERENCES

[1] R. S. Ledley and L. B. Lusted, "Reasoning foundation of medical diagnosis," *Science*, Vol. 130 (1959), pp. 9–21.

[2] J. E. Overall and C. M. Williams, "Models for medical diagnosis," *Behavl. Sci.*, Vol. 6 (1961), pp. 134–141.

[3] H. A. Warner, A. F. Toronto, L. G. Veasey, and R. Stephenson, "Mathematical approach to medical diagnosis. Application to congenital heart disease," *J. Amer. Med. Assoc.*, Vol. 177 (1961), pp. 177–183.

[4] J. E. Overall and C. M. Williams, "Conditional probability program for diagnosis of thyroid function," *J. Amer. Med. Assoc.*, Vol. 183 (1963), pp. 307–313.

[5] R. L. Engle and B. J. Davis, "Medical diagnosis: present, past and future," *Arch. Int. Med.*, Vol. 112 (1963), pp. 512–543.

[6] M. Lipkin, "The likelihood concept in differential diagnosis," *Persp. Biol. Med.*, Vol. 7 (1964), pp. 485–497.

[7] M. F. Collen, L. Rubin, and L. Davis, "Computers in multiphasic screening," *Computers in Biomedical Research*, Chapter 14, Vol. 1 (edited by R. W. Stacy and B. D. Waxman), New York, Academic Press, 1965.

[8] J. Neyman, *First Course in Probability and Statistics,* Chapter 5, New York, Holt, 1950.

[9] M. F. Collen, L. Rubin, J. Neyman, G. B. Dantzig, R. M. Baer, and A. B. Siegelaub, "Automated multiphasic screening and diagnosis," *Amer. J. Pub. Health,* Vol. 54 (1964), pp. 741–750.